OCO-2 Algorithm Theoretical Basis Document

Oxygen-A Band Cloud Screening Algorithm (ABO2)

JPL document number: D-81520

Christopher O'Dell (CWO) and Thomas E. Taylor (TET)
Colorado State University, Fort Collins, CO, USA.

Edited by Annmarie Eldering
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

21-August-2014

Revision History

Revision	Date	Author(s)	Description
1.00	03-Apr-2014	TET	Intial version created.
1.01	17-Apr-2014	TET	Updated ABO2 testing discussion.
1.02	19-May-2014	TET	Implemented CWO comments.
1.03	22-May-2014	TET	Updates to threshold plots and discussion.
1.04	23-May-2014	TET	Updates to micro-windows plot and discussion.
1.05	27-May-2014	TET	Initial delivery for JPL review.
1.06	05-Jun-2014	CWO	Minor changes.
1.07	10-Jul-2014	CWO	Minor changes for JPL document review.
1.07	21-Aug-2014	TET	Recompiled to fix broken bibtex references.

Contents

1 Introduction **3**

2 The OCO-2 ABO2 cloud screening algorithm **4**
2.1 ABO2 retrieval mechanics . . . 4
2.2 ABO2 retrieval first guess . . . 6
2.3 ABO2 retrieval forward model . . . 7
2.3.1 Gas Absorption . . . 7
2.3.2 Rayleigh Scattering . . . 7
2.3.3 Surface Model . . . 7
2.3.4 Radiative Transfer . . . 8
2.3.5 Solar Model . . . 8
2.3.6 Instrument Model . . . 9
2.3.7 Jacobians . . . 10
2.4 ABO2 retrieval cloud flag and thresholds . . . 10
2.4.1 ABO2 surface pressure test . . . 11
2.4.2 ABO2 albedo test . . . 13
2.4.3 ABO2 chi-squared test . . . 15
2.4.4 ABO2 threshold summary . . . 15
2.5 Use of spectral micro-windows. . . . 16
2.6 ABO2 analysis . . . 18
2.6.1 Performance against simulated data . . . 19
2.6.2 Validating against the MODIS cloud mask . . . 20
2.6.3 Verifying optimization of ABO2 thresholds . . . 24

Acknowledgements **26**

A Example ABO2 input files **27**
A.1 OCO-2 retrieval mode . . . 27
A.1.1 CONFIG.MEN . . . 27
A.1.2 aband_options_oco2.dat . . . 28
A.1.3 ms3_options_oco2.dat . . . 30
A.2 GOSAT retrieval mode . . . 32
A.2.1 CONFIG.MEN . . . 32

1 Introduction

The primary science goal of the National Aeronautics and Space Administration's (NASA) Orbiting Carbon Observatory-2 (OCO-2) satellite is to allow for the quantification of global sources and sinks of atmospheric carbon dioxide (CO_2) by providing high spectral resolution near-infrared (IR) measurements of reflected solar light (Crisp et al., 2004). The radiance measurements of backscattered sunlight in the 0.76 μm O_2 A-band the 2.06 μm and 1.61 μm strong and weak CO_2 bands will be made in three science observation modes (nadir, glint and target) at spatial and temporal resolutions much higher than are currently available from other satellite systems (Crisp and Johnson, 2005).

The high spectral resolution of the OCO-2 measurements will provide the necessary sensitivity to the sources and sinks of CO_2 found near the earth's surface, while simultaneously minimizing systematic measurement errors. The high spatial and temporal sampling will provide 8 IFOV footprints per 3 Hz frame, each footprint at $\simeq$1.3 km cross track and $\simeq$2.3 km along-track. The instrument functions only on the day light side of the orbit, yielding approximately 35k soundings per orbit ($\simeq$1 million per day).

A sounding selector tool will be used to determine the exact soundings that will be processed by the computationally expensive Level 2 $X_{\mathsf{CO_2}}$ retrieval algorithm (Mandrake et al., 2013). The L2 algorithm is a Bayesian retrieval that produces estimates of $X_{\mathsf{CO_2}}$, as well as other state variable values, given the high spectral resolution radiances as inputs (O'Dell et al., 2012) and (Crisp et al., 2012). Current computational speeds and budget limitations dictate that approximately 6-7% of the total number of collected soundings will be processed by L2.

One of the largest drivers of error in the retrieval algorithm is scattering due to cloud and aerosol, which adds uncertainties to the radiance measurements, and hence to the retrieved $X_{\mathsf{CO_2}}$. Although atmospheric scattering is accounted for in the L2 forward model, it is sensitive to even thin optical thicknesses and to their vertical distributions and spectral properties. Screening of contaminated scenes is therefore required in order to maximize the data yield.

This document describes in detail the so-called Oxygen-A Band cloud screening algorithm, which will be one two primary cloud screening tools implemented in the operational OCO-2 processing pipeline. The algorithm was introduced and applied to early GOSAT data in (Taylor et al., 2012), with further analysis performed on OCO-2 simulations given in (O'Dell et al., 2012). An overview of the OCO-2 experiment, and an overview of the measurement details (spectral bands, field-of-view, definitions of footprint, etc) are contained in the data user's guide and the L2 ATBD. It is suggested that the reader review those documents to become familiar with the OCO-2 terminology. The reader may also be interested in the companion OCO-2 cloud screener, the so called IMAP-DOAS algorithm (Frankenberg et al., 2005).

2 The OCO-2 ABO2 cloud screening algorithm

The Oxygen-A Band cloud screening algorithm (ABO2 for short) was developed at Colorado State University (CSU) under the Atmospheric Carbon Observations from Space (ACOS) program specifically to address the need for rapid cloud screening of the large volume of OCO-2 scenes. Previous research had demonstrated that high resolution reflectance spectra in the O_2 A-band measured from space potentially would provide adequate information content on absorption and scattering to identify contamination of scenes, e.g., (Stephens and Heidinger, 2000). The absorption by oxygen molecules in this spectral region (approximately 0.759 μm to 0.771 μm) provides a signature in the measured reflectances that can be inverted via standard retrieval techniques to provide estimates of atmospheric parameters such as surface pressure and surface albedo. Using some prior knowledge of the expected values, these parameters can be interpreted to provide information on cloud contamination within the IFOV of the satellite sensor.

The OCO-2 ABO2 algorithm employs a fast Bayesian retrieval to estimate surface pressure and surface albedo from high resolution spectra of the molecular oxygen (O_2) A-band, near 0.765 μm. The radiative transfer forward model (FM) assumes a clear-sky condition, i.e. Rayleigh scattering only, such that differences between the modeled and measured radiances are apparent when the measurement scene contains cloud or aerosol. The estimated surface pressure, surface albedo and the χ^2 goodness-of-fit statistic are used in conjunction with the instrument signal-to-noise ratio (SNR) to flag scenes as cloudy or clear.

The algorithm was introduced and tested using O_2 A-band measurements from the Fourier Transform Spectrometer (FTS) and the Cloud and Aerosol Imager (CAI) aboard the Japanese GOSAT satellite. In that study, the MODIS cloud mask combined with the cirrus reflectance values were used as a truth, as detailed in (Taylor et al., 2012). Studies were also carried out on simulated OCO-2 spectra in (O'Dell et al., 2012).

2.1 ABO2 retrieval mechanics

The ABO2 algorithm is an iterative scheme that attempts to minimize the residual of a simulated and measured spectrum. The simulated spectrum are calculated in clear-sky mode, i.e., it is assumed that no clouds or aerosols are present. Scenes that violate the clear-sky assumption will generally yield large spectral residuals, indicating that the measured radiances are contaminated by cloud and/or optically thick aerosol.

The best-fit spectrum is obtained as outlined in (Rodgers, 2000) via minimization of a cost function that assumes no prior information. The cost function is taken to be the χ^2 of the residuals between the measured and simulated spectra,

$$\chi^2(\mathbf{x}) = \big(\mathbf{y} - \mathbf{F}(\mathbf{x_i})\big)^T \, {\mathbf{S_y}}^{-1} \, \big(\mathbf{y} - \mathbf{F}(\mathbf{x_i})\big), \tag{1}$$

where $\mathbf{y}$ is a vector containing the measured spectrum, $\mathbf{F}(\mathbf{x_i})$ represents the forward model evaluated at state vector $\mathbf{x}_i$ on iteration i, and $\mathbf{S_y}$ is the forward model plus measurement

error covariance matrix. Further discussion of the spectral channels contained in $\mathbf{y}$ are given in Section 2.5.

The retrieval state vector contains five parameters: surface pressure (p_s), the offset to an assumed temperature profile (ΔT), surface albedos at the band beginning and end points (α_1 and α_2, respectively), and a wavelength (dispersion) multiplier (f_λ);

$$\mathbf{x} = \begin{pmatrix} p_s \\ \Delta T \\ \alpha_1 \\ \alpha_2 \\ f_\lambda \end{pmatrix} \tag{2}$$

The dispersion multiplier f_λ is necessary to account for short-term drifts in the instrument's detected frequency and the earth-spacecraft Doppler shift. The surface is assumed to be purely Lambertian, even over water surfaces and the surface albedo is assumed to vary linearly across the band. The $\mathbf{S_y}$ matrix contains only instrument noise and is taken to be diagonal to provide relative weights. No other uncertainties have been included.

Starting from a first-guess state vector $\mathbf{x}_0$, Gauss-Newton iteration is used to update $\mathbf{x}$ from iteration i to iteration i+1;

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \left(\mathbf{K}^T \, {\mathbf{S_y}}^{-1} \, \mathbf{K}\right)^{-1} \mathbf{K}^T \, {\mathbf{S_y}}^{-1} \left(\mathbf{y} - \mathbf{F}(\mathbf{x_i})\right), \tag{3}$$

where $\mathbf{K}$ denotes the Jacobian, defined by

$$\mathbf{K} = \frac{\partial \mathbf{F}(\mathbf{x_i}))}{\partial \mathbf{x}_i}, \tag{4}$$

which are calculated using finite differences.

This simple five-parameter forward model has been shown to be quite linear in state space. Given a reasonably accurate first guess for certain state vector parameters, such as the wavelength multiplier for which the cost function is highly nonlinear, it will converge rapidly. Because calls to the forward model are relatively slow, *only a single update of the state vector based on Eqn. (3) is used in operation.* Additional iterations have been shown to typically yield virtually no change in the state vector.

The code therefore starts at the first-guess $\mathbf{x}_0$ and computes the retrieved state $\mathbf{x}_1$ using a single execution of Eqn. (3). This requires one call to the forward model to evaluate the simulated spectrum, and four additional calls to the forward model to obtain derivatives with respect to p_s, ΔT, α_1 and f_λ. Simple finite differencing is used to calculate the derivatives. The derivative of α_2 is obtained directly via a simple approximation described in Section 2.3.7.

Technically, an additional forward model call at the state $\mathbf{x}_1$ is required in order to evaluate χ^2 at the retrieved state via Eqn. (1). However, assuming linearity of the forward

model between the first-guess and retrieved state, $\mathbf{F}(\mathbf{x}_1)$ can be approximated as;

$$\mathbf{F}(\mathbf{x}_1) \approx \mathbf{F}(\mathbf{x}_0) + \mathbf{K}(\mathbf{x}_1 - \mathbf{x}_0), \tag{5}$$

where the Jacobian matrix $\mathbf{K}$ was defined in Eqn. (4). The advantage is that Eqn. (5) replaces a relatively slow call to the forward model with a fast matrix multiplication and vector addition, using terms that have been calculated previously. Therefore, a total of five calls to the forward model are required for the entire retrieval; one for the simulated spectrum and four more to obtain $\mathbf{K}$. This yields the approximate state $\mathbf{x}_1 \approx \hat{\mathbf{x}}$ that minimizes the χ^2.

On a single desktop CPU, the retrieval run time is on the order of 1 second per sounding for the full GOSAT and OCO-2 spectral grids. However, the run time is largely dependent on the number of frequency points selected. In Section 2.5, the use of spectral micro-windows is employed to speed up the retrieval to approximately 0.3 s per sounding in OCO-2 mode. For GOSAT, the full spectral range is used.

2.2 ABO2 retrieval first guess

The surface pressure and temperature profiles used for the first guess are derived from a linear interpolation in time and space of the standard ECMWF global forecast on 91 vertical layers and one-quarter degree horizontal resolution. The ECMWF product, subset to the OCO-2 orbit will be obtained operationally in the data processing pipeline. The initial guess for ΔT is zero.

The first guess values of surface albedo are evaluated for two frequencies in the continuum near the left and right band edges, $\tilde{\nu}_1 = 12972\,\mathrm{cm}^{-1}$ ($0.77089\,\mu$m) and $\tilde{\nu}_2 = 13188\,\mathrm{cm}^{-1}$($0.75827\,\mu$m). The albedo at each of these points is evaluated directly from the spectrum as

$$\alpha(\tilde{\nu}_i) = \pi\,\frac{\left\langle I(\tilde{\nu}_i, \Delta\tilde{\nu})\right\rangle}{\mu_0\, D_\odot^{-2}\, F_0}, \tag{6}$$

where μ_0 is the cosine of the solar zenith angle, $D_\odot$ is the sun-earth distance in astronomical units (AU) at the time of the observation, and F_0 is the downwelling solar flux density at the top of the atmosphere at the frequency of interest for $D_\odot = 1$ AU. The term $\langle I(\tilde{\nu}_i, \Delta\tilde{\nu})\rangle$ represents an average of the measured intensities from $\tilde{\nu}_i - \Delta\tilde{\nu}$ to $\tilde{\nu}_i + \Delta\tilde{\nu}$. The width $\Delta\tilde{\nu}$ of the average is set to avoid solar and telluric absorption features, and is taken to be $4\,\mathrm{cm}^{-1}$ at $\tilde{\nu}_1$ and $2\,\mathrm{cm}^{-1}$ at $\tilde{\nu}_2$. To account for continuum absorption in the O_2 A-band an atmospheric correction term for Rayleigh scattering should be considered when calculating the first guess. However, it has been neglected in this work and Eqn. 6 is only approximate.

The first guess for the dispersion multiplier ($f_{\lambda,0}$) is derived from the observed position of a known strong solar line in a region devoid of other solar and telluric absorption features. Specifically, we use the solar line at $12985.16325\,\mathrm{cm}^{-1}$ ($0.77011\,\mu$m), and fit for a Gaussian depression relative to the continuum for all channels falling within $3\,\mathrm{cm}^{-1}$ ($\simeq$0.18 nm) of this line. The solar Doppler shift is taken into account in the assumed position of the solar line, as it can cause a shift of approximately $0.02\,\mathrm{cm}^{-1}$ ($\simeq$0.001 nm) and is thus non-negligible.

This approach has an accuracy of order $0.02\,\mathrm{cm}^{-1}$, and is sufficient for the purposes of an accurate first-guess.

Note that the methodology for the ABO2 first guess holds true for both GOSAT and OCO-2 O_2 A-band spectra.

2.3 ABO2 retrieval forward model

The ABO2 retrieval has at its heart the forward model $\mathbf{y} = \mathbf{F}(\mathbf{x})$, where again $\mathbf{y}$ is the vector of radiances (the section of the A-band spectrum to be used), and $\mathbf{x}$ is the 5-parameter state vector. Just like the full L2 retrieval code, the ABO2 forward model contains a number of components: gas absorption optical properties, molecular Rayleigh scattering, a surface model, polarized radiative transfer, a solar model, and an instrument model. We briefly discuss each of these components below.

The vertical layering scheme is very simple in our forward model: only 12 vertical levels are used. They are roughly equally spaced in pressure. The prior profile of temperature and water vapor are interpolated from the automatically-produced ECMWF forecast, and is the same profile as used by the L2 code. We interpolate these to our 12-layer grid.

2.3.1 Gas Absorption

In the O2-A band, we only calculate the absorption due to molecular oxygen; all other species are ignored. We use a look-up table in pressure, temperature, water vapor mixing ratio, and wavenumber that was supplied by the JPL "ABSCO" team, and is the same spectroscopy used in the L2 forward model. The spectral grid spacing is every 0.01 cm^{-1}. The O_2 spectroscopy takes into account line mixing and collisionally-induced absorption, but still assumes a Voigt line shape (though it is recognized that this is inadequate and will eventually be updated to a more realistic model). Typical errors from the spectroscopy are on the order of 1-10 hPa in terms of retrieved surface pressure.

2.3.2 Rayleigh Scattering

Because the Rayleigh scattering optical depth at these frequencies is about 0.02, it cannot be completely ignored as it has a non-neligible effect on the radiative transfer. Again we use the same Rayleigh scattering formulation as for the L2 code. The optical depth in each atmospheric layer is calculated via the model of Bodhaine et al. (1999). Polarization is explicitly taken into account by calculating the full phase matrix appropriate to Rayleigh scattering.

2.3.3 Surface Model

The surface model is assumed to be Lambertian over both land and water surfaces. The Lambertian albedo is assumed to vary linearly across the spectrum (for GOSAT, this is linear in wavenumber; for OCO-2, linear in wavelength). This extreme assumption over

water surfaces is justified because almost no photons will scatter off of both the surface and the Rayleigh atmosphere and make it to our sensor. However, in a future version we may change to using a more traditional Cox-Munk facet model to be more realistic over ocean. We do note that in glint mode, we will regularly retrieve albedos greater than unity, which is perfectly acceptable, as we interpret the albedo to simply be the *effective* surface reflectivity in the sun-surface-satellite scattering geometry.

2.3.4 Radiative Transfer

To have an efficient forward model, the radiative transfer must be as fast as possible. Even for OCO-2, where only 50 cm^{-1} of spectral width is calculated, this is still 5001 and monochromatic points (spaced every 0.01 cm^{-1}). To perform the radiative transfer on each of these would be inordinately slow. Therefore, we adopt the method of low streams interpolation as described in O'Dell (2010), wherein full-resolution (monochromatic) calculations are done for all wavenumber points at a very low accuracy, and a small number (20) set of high-accuracy calculations are performed as well. The LSI method marries these two sets of calculations to obtain a fairly accurate set of monochromatic radiances for stokes parameters I, Q, and U as follows.

In the low-accuracy RT calculations, a simple first order of scattering calculation is performed, for total intensity only. Q and U are not calculated. These low-accuracy calculations are very rapid and are performed for each 0.01 cm^{-1} spectral point.

For the high-accuracy calculations, a 4-stream scalar calculation is performed using the Successive Orders of Interaction (SOI) radiative transfer code (Heidinger et al., 2006; O'Dell et al., 2006), in order to calculate stokes I. Stokes Q and U are calculated using a simple first-order of scattering model. At very high solar zenith angles, we use a 6- or 8-stream model instead, as for these geometries, we find that using only 4 streams can lead to several hPa errors in the retrieved surface pressures. Specifically we use 6 streams for SZA between 77.5° and 82.5°, and 8 streams for SZA greater than 82.5°.

The radiative transfer therefore yields a relatively accurate stokes I, and a first-order-of-scattering estimate of stokes Q and U, for each monochromatic spectral point (every 0.01 cm^{-1}). Note that stokes Q and U arise purely from Rayleigh scattering, as we take the surface to be Lambertian. For ocean glint, this is not true in real life, but the excess photons will simply be attributed to a higher surface albedo.

The code is run in clear sky mode, meaning that the input atmosphere is assumed to be free of all cloud and aerosol and only Rayleigh scattering is considered. Therefore, the direct component of radiation is much larger than the diffuse term, and justifies the simplifying assumptions made above.

2.3.5 Solar Model

The solar model used in the ABO2 forward model is essentially the same as used in the L2 code. It assumes a solar linelist built specifically for OCO-2 by Geoff Toon at JPL. The

linelist contains tens of thousands of solar lines and their parameters for location, width, strength, etc, as derived from both surface and satellite-based solar observations. The sun is assumed to completely fill the OCO-2 and GOSAT Field of View. The solar transmittance is multiplied by the solar continuum, which is taken from a polynomial fit to a standard solar continuum model. The solar doppler shift is also calculated, by internally calculating the sun-earth velocity along the line of sight, as well as the location earth rotation towards or away from the sun.

2.3.6 Instrument Model

The instrument model contains two important components: the spectral model, which contains the spectral response function (ILS) of each channel in the fitted spectral range, and the polarization model, which describes the response of each channel to stokes I, Q, U, and V. Also, we must be aware of "bad pixels" and ignore these in our fits.

The spectral response function of each channel for both GOSAT and OCO-2 are taken from the calibration teams for each instrument. For GOSAT, this is a sinc+boxcar convolution model appropriate to a Fourier Transform spectrometer. For OCO-2, the ILS resembles more of a gaussian but is not exactly so. The ILS model of OCO-1 was originally published in Day et al. (2011). We use the ILS's appropriate to OCO-2.

In GOSAT mode, the code fits to the total intensity I, approximated as the average of the P and S channels on GOSAT, which are orthogonal polarization directions (therefore their average will be nearly identical to total intensity I). In contrast, OCO-2 will measure only a single linear polarization. In nadir and glint modes, the direction of the measured polarization is perpendicular to the principle plane defined by the earth-sun-satellite geometry. The instrument's measured radiance in these viewing geometries is therefore approximated as:

$$I_{\mathrm{meas}} = \frac{\mathrm{I} - \mathrm{Q}}{2}, \tag{7}$$

and so is sensitive to polarization. In target mode, OCO-2 will not generally be oriented like this, so it will be sensitive to both I, Q, and U. The sensitivities to each are called "Stokes coefficients" and are given in the L1B files and are used as inputs to the ABO2 code. Therefore, we use a more general equation inside the code that simply uses these input coefficients:

$$I_{\mathrm{meas}} = m_I \mathrm{I} + m_Q \mathrm{Q} + m_U \mathrm{U} \tag{8}$$

For principle plane viewing, we'll have $m_I = \frac{1}{2}$, $m_Q = -\frac{1}{2}$, and $m_U = 0$.

One more feature of the instrument model is the so-called "bad-pixel map". This is simple a list of channels to be excluded from any analysis, due to bad, spiky or unstable behavior exhibited by these detectors. The bad pixel map will be included in the standard OCO-2 L1B radiance file. The ABO2 reads this list of bad channels and does not include them in any analysis.

2.3.7 Jacobians

As stated previously, the Jacobians are calculated with simple finite differences, as the forward model is not linearized. Because there are only five state vector parameters, this is not a terrible computational burden (though represents a potential way to speed up the algorithm in the future). This generally requires an additional forward model call for each parameter. However, we can get away with only four such calls instead of five, because the derivative with respect to α_2 of the radiance I in a channel at frequency $\tilde{\nu}$ can be computed using the following approximation:

$$\frac{\partial I}{\partial \alpha_2} \approx \frac{\partial I}{\partial \alpha_1} \frac{\tilde{\nu} - \tilde{\nu}_1}{\tilde{\nu}_2 - \tilde{\nu}}. \tag{9}$$

2.4 ABO2 retrieval cloud flag and thresholds

A sanity check of individual soundings is performed against the following criteria prior to running the ABO2 retrieval;

$$\mathcal{F}_A = 2 \quad \text{if} \quad \begin{cases} \text{SNR} < \text{SNR}_{\text{min}} \ \ \text{OR} \ \ \text{SNR} > \text{SNR}_{\text{max}}, \\ \text{SZA} > \text{SZA}_{\text{max}}, \\ f_{\lambda,0} - 1 > f_{\lambda,\text{max}}, \end{cases} \tag{10}$$

where SNR is the measured signal-to-noise ratio, SZA is the solar zenith angle and $f_{\lambda,0}$ is the retrieval first guess wavelength multiplier. The current operational values of SNR_{min}, SNR_{max}, SZA_{max} and $f_{\lambda,\text{max}}$ are 20.0, 10^4, 85.0 and 0.2, respectively.

The lower bound of the SNR check removes noisy scenes for which the retrieval will be unable to produce a meaningful fit, while the upper bound removes scenes that likely have invalid radiance values at one or more spectral channels. The check on SZA removes scenes when the sun is near the horizon - cases for which the radiative transfer is likely to fail due to invalid assumptions (such as the plane-parallel approximation). The L2 code has a similar SZA screening, so it makes sense to remove the scenes at this step in the processing. The check against $f_{\lambda,\text{max}}$ is to remove cases for which the code failed to find the solar line at 12985 cm^{-1}. Such failure usually indicates that there is a problem with the O_2 A-band spectrum. Scenes that meet any of the above conditions are classified as undetermined and the ABO2 cloud flag ($\mathcal{F}_A$) is set to 2.

All soundings with $\mathcal{F}_A \neq 2$, are then processed thru the ABO2 retrieval described in the preceding sections. The ABO2 cloud flag $\mathcal{F}_A$ is set to 1, indicating cloud, if any of the

following three conditions occur:

$$\mathcal{F}_A = 1 \quad \text{if} \quad \begin{cases} |\Delta p_{s,cld}| > \Delta p_{s,t} \\ \overline{\alpha} < \alpha_t^{\text{LO}} \ \text{ OR } \ \overline{\alpha} > \alpha_t^{\text{HI}} \quad , \quad \text{else} \quad \mathcal{F}_A = 0. \\ {\chi^2}_r > {\chi^2}_t(\text{SNR}) \end{cases} \tag{11}$$

The value $\Delta p_{s,cld}$ in line 1 of Eqn. (11) (referred to as "dp cloud") is calculated as;

$$\Delta p_{s,cld} = p_s - p_{s,ap} - p_{s,0}(\text{SZA}), \tag{12}$$

where p_s is the retrieved surface pressure, $p_{s,ap}$ is the prior surface pressure derived from the ECMWF model and $p_{s,0}(\text{SNR})$ is an empirically determined clear-sky surface pressure offset that is dependent on solar zenith angle. The value $\Delta p_{s,t}$ in Eqn. (11) is the surface pressure difference threshold, the determination of which is discussed in Section 2.4.1.

Also in Eqn. (11), the value $\overline{\alpha}$ is the average of the retrieved albedos at the band end points, α_t^{LO} and α_t^{HI} are low and high albedo thresholds, ${\chi^2}_t(\text{SNR})$ is the chi-squared goodness-of-fit threshold and ${\chi^2}_r$ is the reduced χ^2 value, defined as;

$${\chi^2}_r = \frac{\chi^2}{m-n} \tag{13}$$

where m is the number of fitted channels and $n = 5$ is the number of retrieval variables. Detailed discussions of the surface albedo and chi-squared parameters are given in Section 2.4.2 and Section 2.4.3, respectively.

The ABO2 algorithm is clear sky conservative in that if any combination of the tests described in Eqn. (11) are meet, then the cloud flag is set to 1, prescribing the scene as cloudy. Otherwise, the scene is regarded as clear and $\mathcal{F}_A = 0$. Each sounding is classified as either land or water using the land fraction from the L1B files, where scenes with land fraction less than 0.20 are considered to be water surfaces. Note that the ABO2 algorithm provides a simple integer cloud flag for each OCO-2 sounding, i.e., no fractional clear confidence level as with some cloud screening algorithms, e.g., (Ishida and Nakajima, 2009).

Ideally, screening variables should have no dependencies on intrinsic quantities like SZA or SNR. Unfortunately, there are unknown as well as known sources of bias, such as spectroscopic error and instrument calibration. Optimal selection of the thresholds was achieved using a set of real GOSAT data consisting of ≃19,000 post-filtered soundings returned from the ACOS B2.9 L2 retrieval. These data served as a proxy for clear-sky scenes and were subdivided into ≃11000, 2300 and 5600 soundings for land H-gain, land M-gain and water scenes, respectively. A similar analysis will be employed for OCO-2 once real data is available.

2.4.1 ABO2 surface pressure test

The test on the retrieved surface pressure (p_s) has been shown to be the most powerful of the three ABO2 checks. The value of "dp cloud" ($\Delta p_{s,cld}$) given in Eqn. (12) requires an

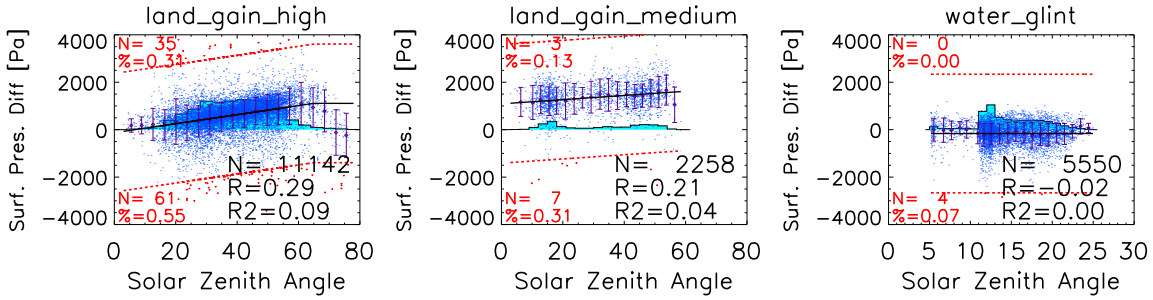

Figure 1: Surface pressure difference ($p_s - p_{s,ap}$) versus SZA for land h-gain (left), land m-gain (middle) and glint (right). Individual data points are shown as blue dots with binned average values in purple with vertical bars representing one-sigma standard deviations. The least squares linear fit (black solid lines) give the bias ($p_{s,0}$(SZA)). Red, dashed lines indicate the filtering threshold ($\Delta p_{s,t}$) with the number of rejected scenes listed.

accurate a priori value ($p_{s,ap}$), as well as the clear-sky bias value ($p_{s,0}$(SZA)). The ECMWF model used for the a priori has been shown to be accurate to $\simeq$2 hPa at most times and locations (Salstein et al., 2008). The data is provided in near real time and is ingested as part of the JPL processing pipeline for both GOSAT and OCO-2 operation.

Under true clear-sky conditions, when very little photon path length modification takes place, the retrieved surface pressure from high resolution O_2 A-band measurements should agree with the ECMWF forecast value to within a few hecta Pascals (hPa). Any bias is believed to be dominated by spectroscopic effects, although instrument mischaracterization could also play a role. The ACOS L2 B2.9 test set was used to determine the clear-sky surface pressure offset as a function of SZA, $p_{s,0}$(SZA) in Eqn. (12). As shown in Figure 1, a weak correlation exists between Δp_s and SZA. A piece-wise linear fit in SZA is used to determine the values of $p_{s,0}$(SZA), which are stored in an auxiliary ABO2 input file. Currently the "knees" of the fit for scenes measured over water are at SNR=20 and SNR=70. This is done separately for high gain land, medium gain land and water glint scenes. Analysis on the GOSAT test set indicates that less than 1% of the clear-sky soundings are rejected using the current settings.

For scenes contaminated by cloud or thick aerosol, there will in general be an increase in the photon path length distribution due to multiple scattering. This will manifest itself as higher than expected retrieved p_s, hence large discrepancies with the ECMWF value (i.e., large $\Delta p_{s,cld}$). For scattering layers residing high in the atmosphere (e.g. cirrus clouds), photon path length shortening occurs due to reflection back to space prior to penetration of the full atmosphere profile. This will yield lower than expected p_s and again result in large values of $\Delta p_{s,cld}$ of the opposite sign. Both path lengthening and shortening are flagged by implementation of the two-sided filter shown in Eqn. (11).

The current setting of Δp_t is 25.0 hPa for all land scenes and water scenes when SNR $>$ 70.0. This applies to virtually all GOSAT soundings and should be true for OCO-2, except for measurements over water in nadir viewing mode (which will occur frequently in alternating 16 day repeat cycles). The 25.0 hPa setting is relatively modest given the

typical ECMWF p_s accuracy of a few hPa referenced earlier. Scenes must be fairly cloudy in order to be flagged, while scenes with very thin scattering layers will likely pass. See the detailed discussion of algorithm performance in Section 2.6.

Because water absorbs strongly in the 0.76 μm O_2 A-band, the SNR decreases rapidly as the instrument's field of view (FOV) moves away from the specular glint spot. This makes accurate retrievals of surface pressure more difficult. Therefore, for soundings measured over water surfaces, the reliance on p_s is relaxed via increasing the value of the threshold, i.e., $\uparrow\Delta p_t$ with $\downarrow$SNR, This forces the requirement for a larger difference between retrieved and prior surface pressure in order to flag a scene as cloudy. For water scenes with SNR between 70.0 and 20.0, Δp_t is set to 50.0 hPa. For SNR values < 20.0, Δp_t is set to 100.0 hPa.

The at-launch settings for OCO-2 are the same as those described above. Once real data is available, a complete analysis will take place to update the thresholds.

2.4.2 ABO2 albedo test

For land viewing scenes, the surface albedo check is difficult to implement due to the high spatial variation in the reflectivity of the earth's surface. Although global ground cover vegetation maps exist, surfaces are often covered with snow and or change with seasons, etc. Therefore the ABO2 essentially uses the albedo test as a simple sanity check over land by setting the values of α_t^{LO} and α_t^{HI} to 0.0 and 1.0, respectively.

The situation is different over water viewing surfaces, where the surface reflectivity is a reasonably well behaved function of sun and satellite viewing geometries and to a certain extent the wind speed which produce foam white caps. Threshold values of α are determined based on the sun and satellite geometries, with scenes subdivided into three bins as follows;

(1) $\Theta_g > 30^\circ$ (non-glint),

(2) $3^\circ < \Theta_g < 30^\circ$ (partial-glint),

(3) $\Theta_g < 3^\circ$ (full glint),

where Θ_g gives the angular difference in degrees between the center of the sounding IFOV and the point of solar specular reflection, and is called the "glint angle".

For the non-glint case, there is high contrast between bright cloud and the dark water surface, allowing the ABO2 algorithm to rely primarily on the retrieved surface albedo. The α_t^{HI} threshold is therefore set to 0.05 when the glint angle (Θ_g) is larger than 30°.

For the partial-glint viewing cases, an empirically determined value of the surface albedo threshold that depends on Θ_g was determined for GOSAT data as;

$$\alpha_t^{\mathrm{HI}} = 0.2 - 0.15/27.0\,(\Theta_g - 3.0). \tag{14}$$

For OCO-2 prelaunch (where no data is available for analysis) the value of α_t^{HI} is set to 10.0. This will likely be updated soon after launch when real data can be used to make an

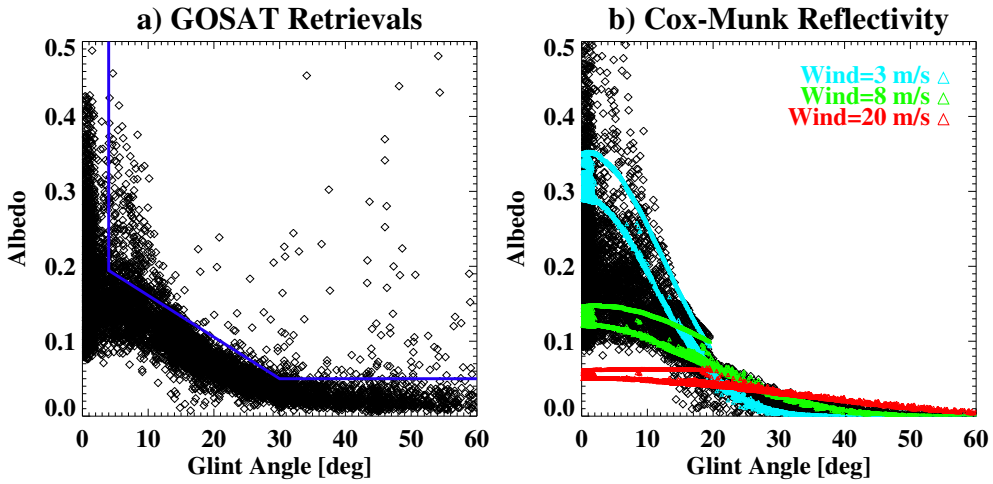

Figure 2: (a) Retrieved O_2 albedo for scenes passing the IMAP, $\Delta p_{s,cld}$, and χ^2 tests, which means these scenes are mostly clear. The blue line represents the value below which a scene must be to pass the albedo test over ocean. It is purely empirical, and is meant to encompass the bulk of the scenes. (b) Theoretical reflectively calculated for an isotropic Cox-Munk facet model (Cox and Munk, 1954a,b), using the observed geometry for each scene in panel (a). Three different fixed wind speeds are calculated, as well as the estimated wind speed for each scene from the ECMWF forecast.

empirical fit as was done for GOSAT.

For the full-glint case, the value of $\alpha_t^{\rm HI}$ is increased to 1000.0, since the contrast between cloud and water is not as apparent. This essentially disables the albedo check.

Figure 2 provides a graphical representation of the albedo test over ocean. Panel (a) shows the retrieved O_2 albedo for a set of scenes passing the $\Delta p_{s,cld}$ (surface pressure) test, a weak χ^2 test (described in the next section), and the IMAP CO_2 and H_2O ratio tests, for a large set of GOSAT over-ocean observations. The blue line represents the albedo threshold employed over ocean; points above this line fail the albedo test and are labelled cloudy. This line is purely empirical. A better test might be to use a theoretical estimate of what the retrieved surface albedo should be under clear sky conditions, which will depend on the primarily wind speed and viewing geometry. Figure 2b shows a simple model of this for three specific wind speeds (3, 8, and 20 m/s) in the colored symbols, and using the estimated surface wind speed based on the ECMWF forecast (black symbols). The main deficiency is that for non-glint viewing conditions (glint angle $\gtrsim 30°$), the observed albedos tend to be high than the pure Cox-Munk reflectivities. However, this is likely easily accounted for via a somewhat higher threshold than the theoretical value. Therefore, in the future we'll likely update this test to have a more physical based threshold.

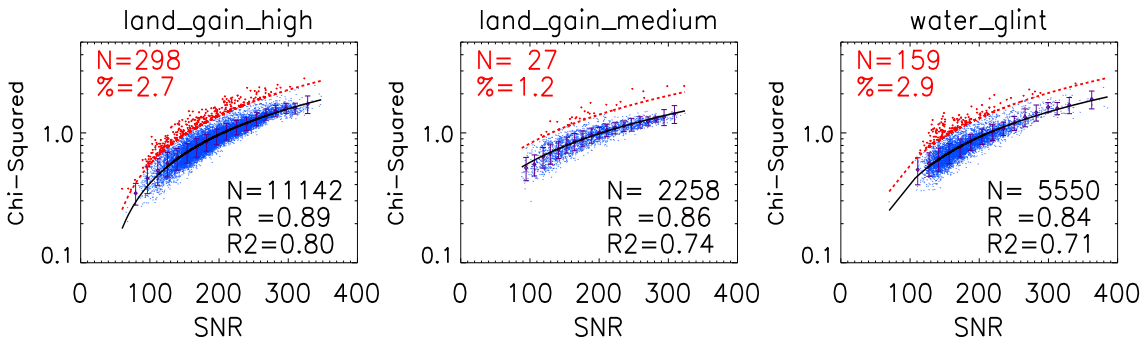

Figure 3: χ^2 versus SNR for land h-gain (left), land m-gain (middle) and glint (right). Individual data points are shown as blue dots with binned average values in purple with vertical bars representing one-sigma standard deviations. The least squares linear fit is shown as a solid black line. Red, dashed lines indicate the filtering threshold (${\chi^2}_t$(SNR)) with the number of rejected scenes listed.

2.4.3 ABO2 chi-squared test

For clear-sky conditions, where the goodness-of-fit is limited by instrument noise, the ABO2 retrieval should produce chi-squared values (χ^2) near unity. However, this condition is rarely satisfied because the fit is limited by uncertainties in spectroscopy and shortcomings in the instrument calibration, rather than the instrument signal to noise ratio (SNR). For clear-sky conditions, we find that χ^2 scales exponentially with increasing SNR. The correlation breaks down in the presence of clouds, which introduce optical path length errors that produce much larger values of χ^2.

An empirical fit was made to the GOSAT test data with the functional form;

$$\chi^2 = a \exp^{b \cdot \mathrm{SNR}}, \tag{15}$$

where a and b are determined by a linear fit of $\ln(\chi^2)$ to SNR. Due to the sensitivity of the noise properties to both the surface albedo and the gain setting, three sets of fitting coefficients were defined, for the land high gain, land medium gain and water surface cases, independently.

The strong correlation between χ^2 and SNR is depicted in Figure 3. A constant scale factor (SF) was determined based on the linear fit to avoid screening the set of scenes with χ^2 slightly higher than the fit value. That is;

$${\chi^2}_t(\mathrm{SNR}) = \mathrm{SF} \cdot {\chi^2}_{r,0} \tag{16}$$

Currently SF=1.4, allowing more than 97% of the clear-sky soundings in the GOSAT test set to pass.

2.4.4 ABO2 threshold summary

It should be noted that, the selection of the thresholds for the ABO2 algorithm was based on a somewhat limited data set. Therefore, it is possible that we have introduced biases

Description	Spectral Range (cm^{-1})	Spectral Range (μm)	Num Channels (m)
Full GOSAT O_2 A-band range	12869-13330	0.750-0.777	2310
Truncated GOSAT range	12968-13190	0.758-0.771	1113
Full OCO-2 O_2 A-band range	12968-13190	0.758-0.771	1016
OCO-2 Micro 1	13145-13172	0.759-0.761	89
OCO-2 Micro 2	13047-13072	0.765-0.766	83

into the results. Future research should expand the test data set. For OCO-2 prelaunch, the empirically determined GOSAT data will be employed until OCO-2 specific values can be derived.

For completeness, example input files are shown in section A for both the OCO-2 and GOSAT retrieval modes.

2.5 Use of spectral micro-windows.

The measurement vector $\mathbf{y}$ used in the ABO2 retrieval (Refer to Eqn. (1) in Section 2.1) contains the set of frequency-dependent O_2 A-band radiance values. Although the full spectral range of channels can be used, the information contained in the closely spaced channels are not completely independent i.e., do not each provide independent pieces of information to constrain the retrieval. Since the algorithm run-time is proportional to the number of channels contained in $\mathbf{y}$ ($m =$ number of channels), a decrease in m will provide a decrease in computation time.

For GOSAT data processing, which has a manageable data volume, the ABO2 run-time is not so critical. Therefore, the set of spectral channels that overlap the OCO-2 O_2 A-band region are used in $\mathbf{y}$. The spectral ranges and number of channels are summarized in Table 2.5. The algorithm runs at approximately 1 second per sounding with this configuration.

However, for OCO-2 processing, which will have a data volume approximately 100 times that of GOSAT, an effort was made to increase the computational efficiency of the ABO2 algorithm. This was achieved by intelligently selecting truncated frequency ranges (called "micro-windows") from the full frequency grid to reduce m, while still providing accurate retrieval of p_s and α.

It was determined that the combination of a set of two micro-windows, each of approximately 0.002 μm width (2 nm), provided suitable accuracy while reducing the ABO2 run time to about 0.3 seconds per sounding. Micro-window 1 stretches from the edge of the continuum region into the strong absorption feature of the O_2 A-band. Micro-window 2 spans three O_2 doublet features in the P-branch of the O_2 A-band. Figure 4 shows the micro-window spectral ranges for reference. The new ABO2 run speed falls just under the upper limit of the mission requirement specified by the data processing team at JPL.

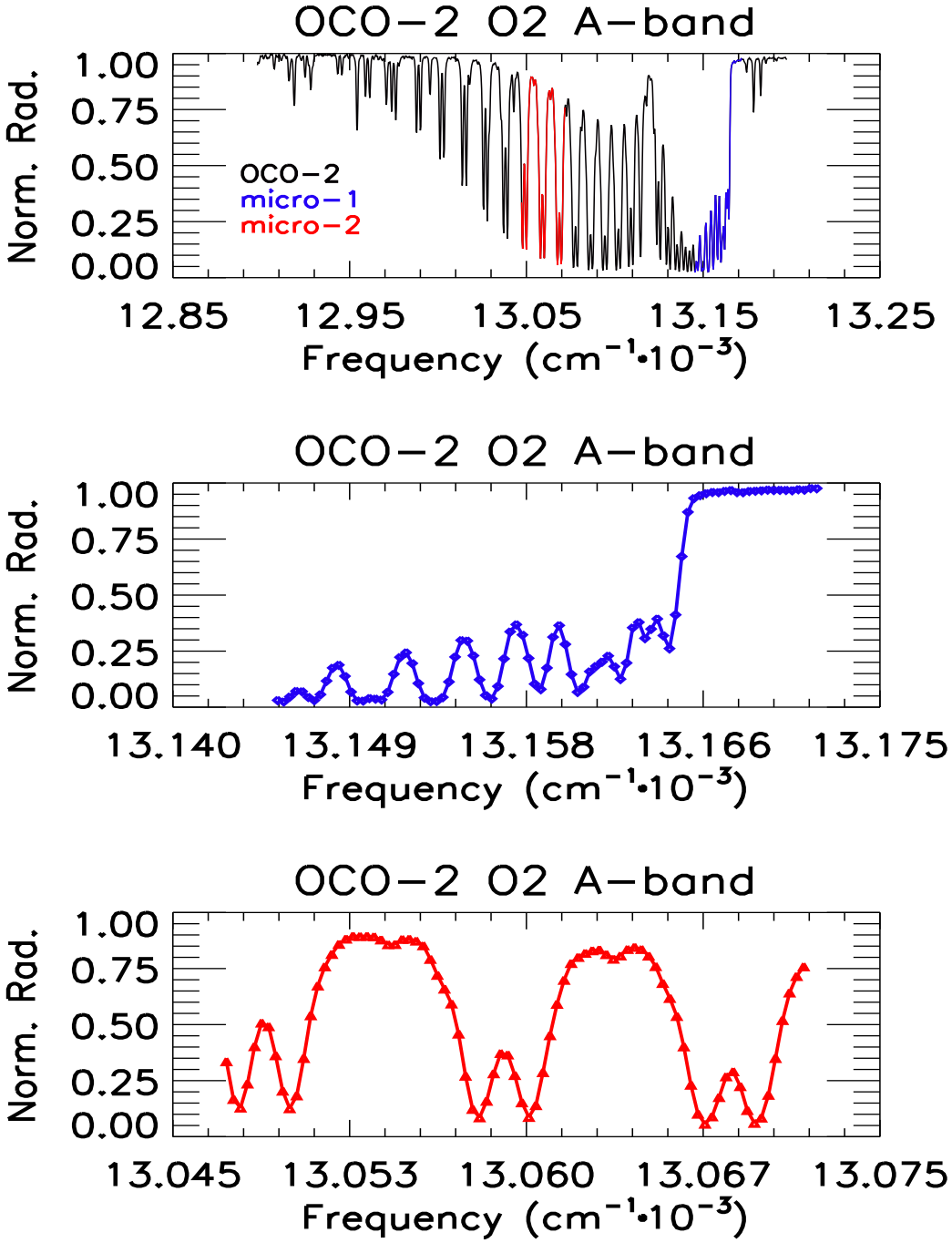

Figure 4: OCO-2 O_2 A-band spectrum showing the two spectral micro-windows used in the ABO2 retrieval $\mathbf{y}$ measurement vector. The top panel shows a typical simulated spectrum that was calculated at 21.8° S latitude and 29.4° E longitude (Zimbabwe, Africa) with a solar zenith angle of 40°. The middle and lower panels show details of micro-window 1 and 2, respectively.

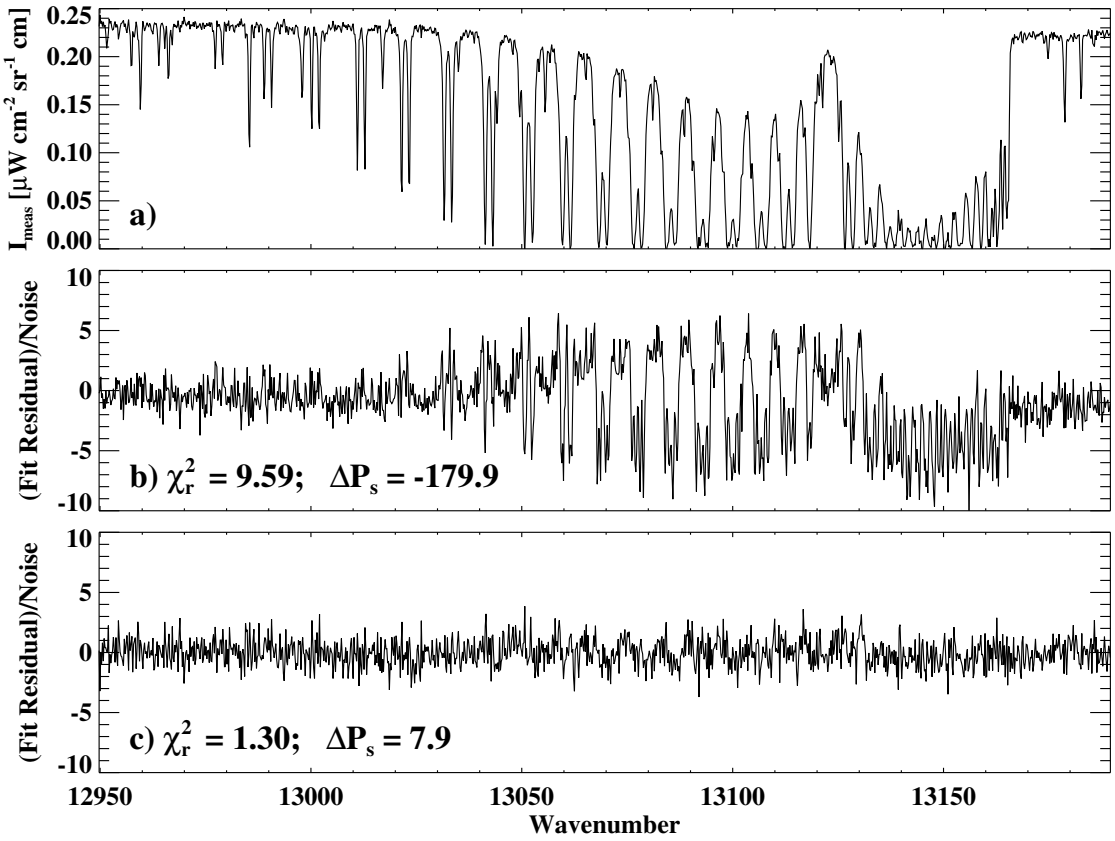

Figure 5: The upper panel shows a typical O_2 A-band spectrum. The middle and lower panels show the residuals for sample cloudy and clear scenes, respectively.

2.6 ABO2 analysis

It has been deduced from theoretical studies on a simulated data set that ABO2 has extremely high accuracy in identifying any scattering material that has optical thickness greater than about 0.2, and that is located approximately 200 hPa or higher above the surface. The upper panel of Figure 5 shows a typical high resolution O_2 A-band spectrum. The middle panel shows the spectral residual for a typical cloudy scene, while the lower panel is the residual for a clear scene. In this example, the cloudy scene violates the p_s and χ^2 threshold tests given in Eqn. (11). For the clear scene, not only is ${\chi^2}_r$ close to unity, but the difference between the estimated surface pressure and the prior value from ECMWF ($\Delta p_{s,cld}$) is small.

Below we give evaluations of the ABO2 algorithm versus simulated OCO-2 data and a comparison to MODIS for GOSAT data. These are synopsis of work first reported in (O'Dell et al., 2012) and (Taylor et al., 2012), respectively.

2.6.1 Performance against simulated data

Tests of the ABO2 algorithm in OCO-2 mode were performed on simulated observations, as detailed in (O'Dell et al., 2012). Whatever its deficiencies, the use of simulated data has the advantage that the "true" values of instrument characteristics, spectroscopy, meteorology, surface properties, clouds and aerosols and other factors are, by definition, known.

The CSU "OCO Simulator" (O'Brien et al., 2009) was used to generate a set of simulated data for ten A-Train orbits spanning 14-16 September, 2006. The scenes contained AOD ranging from near zero to a maximum of around 10, as this is the measurement limit of the CALIPSO product, which was used to generate the synthetic atmospheres. Realistic instrument noise was added to the spectra. Only measurements made over land surfaces were used in this study, yielding 6522 soundings for analysis. The ABO2 cloud screen was run using relatively loose threshold criteria of $\Delta p_{s,t} = 40\,\mathrm{hPa}$ and χ^2 scale factor (SF in Eqn. (16)) $= 2.3$. The ABO2 results were then analyzed using the "true" AOD to define clear versus cloudy conditions.

The overall performance numbers on the simulated data set are summarized in Figure 6. Panel (a) shows the distribution of cloud plus aerosol optical depth (AOD) at $0.76\,\mu$m in the synthetic data set, as well as the fraction of scenes identified as clear. Previous studies have identified 0.3 as a reasonable AOD threshold below which to attempt XCO_2 retrievals (Crisp et al., 2004). Therefore, clear-sky was defined as AOD ≤ 0.3 and cloudy as AOD > 0.3. With this definition, 26% of all scenes are clear, and about 87% of all scenes are classified correctly. However, about one third of scenes that pass the cloud filter are false positives; they were classified as clear but have true AOD > 0.3.

The cloud screening performance is dramatically different for low cloud versus high cloud cases. Figure 6b (Figure 6c) shows the histograms of AOD for high (low) cloud or aerosol cases, in which 95% of the AOD resides in the top 40% (bottom 30%) of the atmosphere. The high cloud cases have been considered by many authors to be the most problematic e.g., (O'Brien and Rayner, 2002) and (Aben et al., 2007), however it is seen that the cloud screening performance here is reasonably good. Virtually all high cloud cases with AOD > 0.3 are classified as cloudy, and virtually all cases with AOD < 0.1 are classified as clear. By contrast, almost all low cloud cases with AOD < 1 are classified as clear, as are more than half of cases with AOD > 1. Most of these are water cloud cases, and they occur disproportionately at higher solar zenith angles.

There may be problems with ABO2 in mountainous regions due to reliance on the ECMWF surface pressure, which is interpolated in time and space. Over rugged terrain this interpolation is certainly more prone to error. Furthermore, over high terrain, clouds are more likely to exist near the surface - a condition leading to failure of the ABO2 due to the minimal change in the photon path length.

Both thin high clouds and thicker low clouds cause problems for the XCO_2 retrieval. This problem can be partially mitigated by simply tightening the surface pressure thresh-

old to $\Delta p_t < 10\,\mathrm{hPa}$, as shown in the dashed lines in Figure 6. This reduces the rate of false positives from one in three to about one in five. In fact, this is typically used as a post-processing filter for ACOS L2 retrievals.

As demonstrated in Figure 7, the small population of low clouds with total optical depth (TOD) around 2.0 that is missed by the ABO2 cloud screener can be successfully identified using the OCO-2 IMAP-DOAS algorithm that was introduced in Frankenberg et al. (2005). The plan for OCO-2 operational processing is to use a combination of retrieval information from the ABO2 and IMAP-DOAS codes as input to the sounding selector algorithm to determine which scenes will be processed by the computationally expensive L2 retrieval algorithm (Mandrake et al., 2013).

2.6.2 Validating against the MODIS cloud mask

Results from the ABO2 algorithm on the GOSAT test set were compared against a cloud mask designed using the MODIS cloud mask, combined with the MODIS cirrus reflectance product. The details of the test are given in (Taylor et al., 2012).

The MODIS cloud mask algorithm employs a series of spectral tests, coupled with pre-set thresholds, to distinguish clear from cloudy scenes (Ackerman et al., 1998, 2006; Frey et al., 2008). Results are reported on a 1 km grid spacing, providing roughly 100 pixels per GOSAT TANSO-FTS footprint.

The first two bits of the cloud mask summarize the overall clear confidence level as

$$\mathcal{F}_M = \begin{cases} 0, \text{confident cloud} & \text{if } 0.00 \geq Q_M < 0.66 \\ 1, \text{probably cloud} & \text{if } 0.66 \geq Q_M < 0.95 \\ 2, \text{probably clear} & \text{if } 0.95 \geq Q_M < 0.99 \\ 3, \text{confident clear} & \text{if } 0.99 \geq Q_M \geq 1.0 \end{cases} . \tag{17}$$

The cirrus reflectance product is derived from reflectance measurements in the $1.38\,\mu$m channel, where water vapor attenuates the signal when present (Gao et al., 1993), (Gao et al., 2002), (Meyer et al., 2007).

The TANSO-FTS soundings are classified as either clear or cloudy using the MODIS cloud mask (MYD35) and the cirrus reflectance contained in the cloud properties (MYD06) product. The MODIS clear sky ratio (CSR_M) and the MODIS mean cirrus reflectance (MCR_M), calculated for each TANSO-FTS sounding as follows;

$$CSR_M = N_{clear}/N_M \tag{18}$$

and

$$MCR_M = \frac{\sum_{i=1}^{N_M} R_{i,cirrus}}{N_M}, \tag{19}$$

respectively. Here, N_M is the total number of MODIS pixels falling within a TANSO-FTS footprint and N_{clear} is the sum of all pixels labeled as either "confident clear" ($\mathcal{F}_M = 3$) or "probably clear" ($\mathcal{F}_M = 2$) in the MYD35 cloud mask product. The value of the cirrus

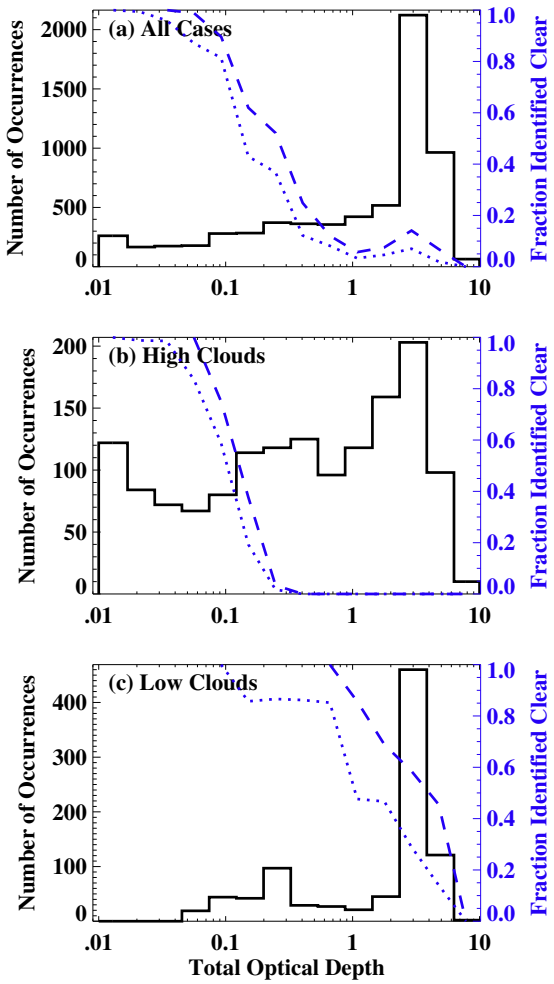

Figure 6: Panel (a): Distribution of cloud plus aerosol optical depth (AOD) for the synthetic orbits (solid line) and fraction of scenes identified as clear by the cloud screening algorithm (dashed and dotted lines). The dashed line shows the result for the operational thresholds, while the dotted line shows the result when the Δp_t threshold is tightened to 10 hPa. Panel (b): Same as panel (a), but only shows those cases where 95% of the AOD resides in the upper 40% of the atmosphere (high clouds). Panel (c): Same as panel (a), but only shows those cases where 95% of the AOD resides in the lowest 30% of the atmosphere (low clouds).

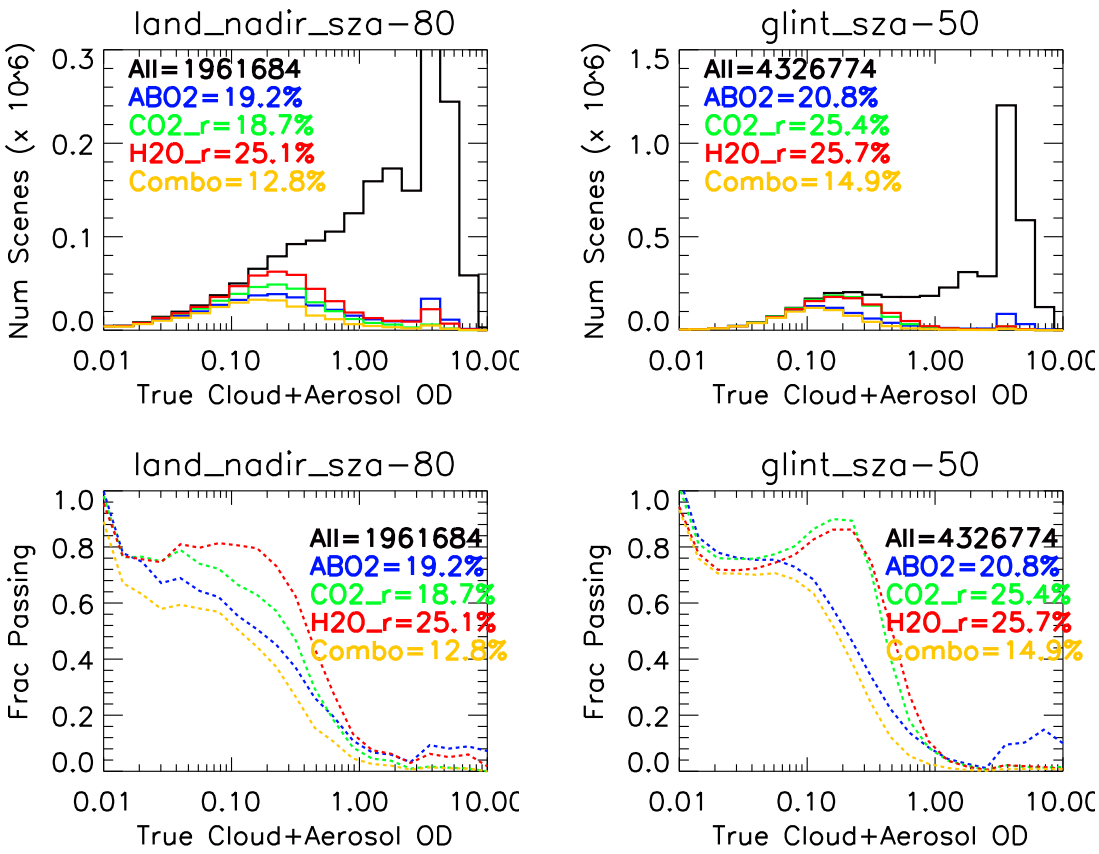

Figure 7: Top panels; Distribution of cloud plus aerosol optical depth (AOD) for land scenes with SZA $< 80°$(left) and for water glint scenes with SZA $< 50°$(right). Multi colored lines indicate all scenes (black), scenes flagged by ABO2 only (blue), scenes flagged by the IMAP-DOAS CO_2 ratio only (green), scenes flagged by the IMAP-DOAS H_2O ratio only (red) and those flagged by the combination of all three (gold). Bottom panels; the fraction of scenes passing as clear.

reflectance reported in the MYD06 cloud properties file is given by $R_{\imath,cirrus}$, where $\imath$ indicates the MODIS pixel index.

$$\widetilde{\mathcal{F}}_M^{(2)} = \begin{cases} 0, \text{ clear} & \text{if } CSR_M > 0.80 \text{ and } MCR_M \leq 0.01 \\ 1, \text{ cloud} & \text{all other scenes} \end{cases} . \tag{20}$$

A statistical comparison of the results from the ABO2 algorithm to the MODIS baseline results was performed via contingency table analysis. The contingency index $\mathcal{I}$ is defined for ABO2 as:

$$\mathcal{I} = \begin{cases} 1, \text{ agreed clear,} & \text{true positive } (TP), \\ 2, \text{ MODIS cloudy, ABO2/CAI clear,} & \text{false positive } (FP), \\ 3, \text{ MODIS clear, ABO2/CAI cloudy,} & \text{false negative } (FN), \\ 4, \text{ agreed cloudy,} & \text{true negative } (TN). \end{cases} \tag{21}$$

From the set of contingency indices, the contingency rates are calculated as,

$$\begin{array}{ll} \text{true positive rate,} & TPR = N_{TP}/P, \\ \text{false positive rate,} & FPR = N_{FP}/N, \\ \text{false negative rate,} & FNR = N_{FN}/P, \\ \text{true negative rate,} & TNR = N_{TN}/N. \end{array} \tag{22}$$

Here, N_{TP}, N_{FP}, N_{FN} and N_{TN} are the numbers of TP, FP, FN and TN cases, respectively, and P and N are the total numbers of clear and cloudy atmospheres, respectively. Note that N should not be confused with the previously defined values of N_M, which represents the number of MODIS pixels falling within a TANSO-FTS footprint. In general, a cloud screening algorithm needs to provide high TPR, while simultaneously minimizing FPR to be useful.

Two diagnostic quantities calculated from the contingency values are the accuracy,

$$ACC = (N_{TP} + N_{TN})/(P + N), \tag{23}$$

and the positive predictive value,

$$PPV = N_{TP}/(N_{TP} + N_{FP}). \tag{24}$$

In this work, the accuracy[1] is interpreted as the skill of the algorithm in successfully classifying scenes as cloudy or clear, relative to MODIS. The positive predictive value[2] gives the percentage of soundings that were identified as clear that are in fact clear.

[1]The accuracy provides the most succinct and easily interpreted description of the algorithm's skill when the number of clear cases is approximately equal to the number of cloudy cases. As the frequency of the cloudy to clear scenes shifts, the interpretation of ACC becomes more difficult and less useful.

[2]As with ACC, some care has to be taken in the interpretation of the PPV because it depends on the relative numbers of clear and cloudy scenes. In this research the value of PPV is in some ways the most critical, as $1 - PPV$ indicates the percentage of soundings passed to the X_{CO_2} retrieval algorithm that are contaminated by cloud.

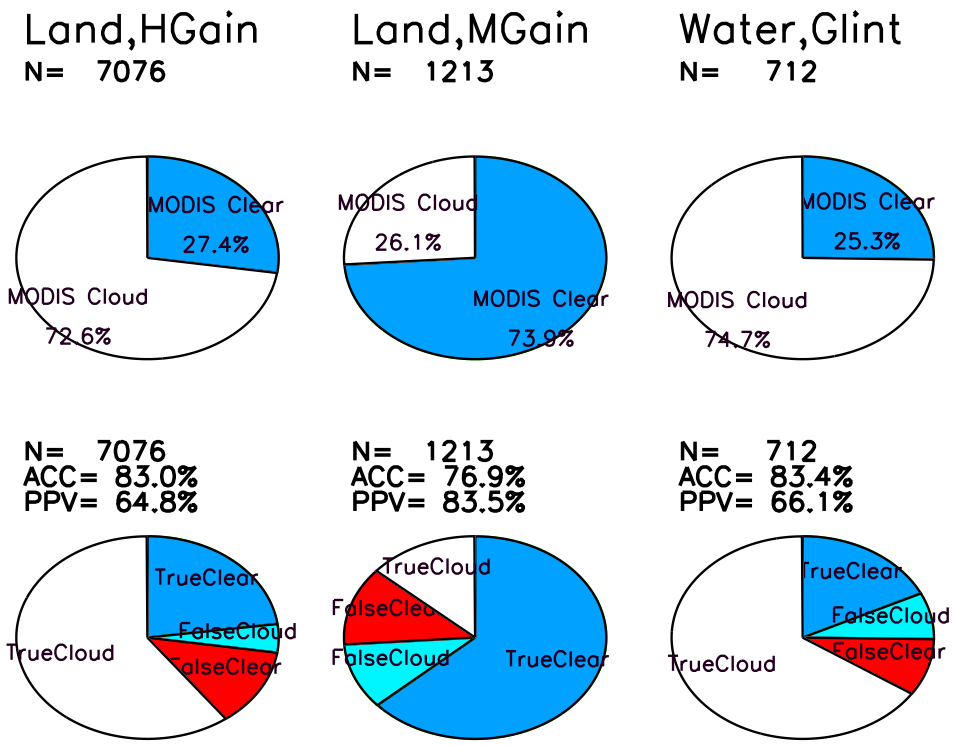

Figure 8: Contingency table comparison of the ABO2 cloud screen to MODIS.

The analysis on the test GOSAT data set is displayed in Figure 8. The accuracy is given by the sum of white and blue pie slices and is 83%, 77% and 83% for land H-gain, land M-gain and glint, respectively. The ABO2 misses $\simeq$15%, 42% and 12% of the cloudy scenes (red slices), while about 20-30% of the clear scenes are incorrectly flaged as cloudy (turquoise slices). The positive predictive values are 65%, 84% and 66% for land H-gain, land M-gain and glint, respectively.

2.6.3 Verifying optimization of ABO2 thresholds

A cloud flagging algorithm that relies on threshold testing should not be overly sensitive to the selection of the values. Here we demonstrate the sensitivity of the ABO2 threshold parameters by performing tests over an ensemble of values and displaying the calculated accuracies as contour maps as shown in Figure 9. In these plots, the white diamonds indicates the selected operational thresholds. Since the contour gradients are not strong near the operational values, it verifies that the accuracy of ABO2 is not overly sensitive to small changes in the selected threshold values.

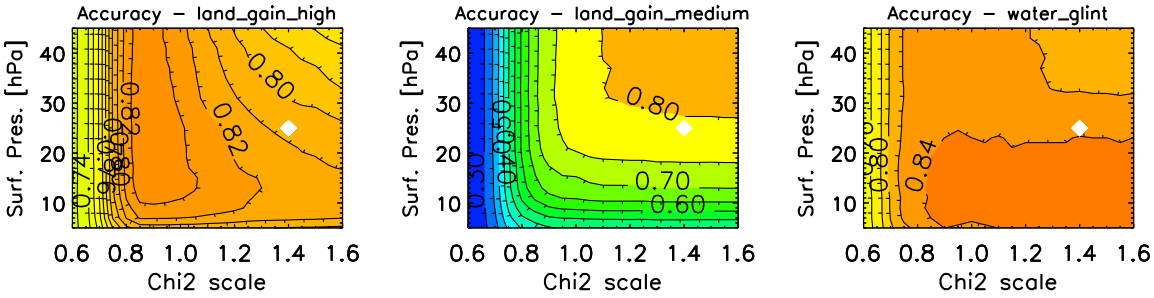

Figure 9: Contour plots of the cloud screen accuracy rates compared to MODIS as a function of Δp_s and χ^2 for land H-gain (left), land M-gain (middle) and glint (right). Current threshold setting represented by white diamond.

Acknowledgments

The research described in this document was funded through the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

A Example ABO2 input files

A.1 OCO-2 retrieval mode

A.1.1 CONFIG.MEN

```
#
# Configuration file for the preprocessor.
# Blank lines and lines beginning with # are ignored.
#
L1B_FILE_PATH = /path1/path2/oco2_L1bScND_example.h5
MET_FILE_PATH = /path1/path2/oco2_ECMWFND_example.h5
QUICK_LOOK_FILE_NAME_OUT =/path1/path2/oco2_L2ABPND_example.h5
SUCCESS_SEMAPHORE_FILE_NAME_OUT = ./semaphore.txt

#
# A-Band CloudScreen Options
#
ENABLE_ABO2_CLOUD_SCREEN = 1
ABO2_OPTIONS_FILE_PATH = aband_options_files/aband_options_OCO2_example.dat
MS3_OPTIONS_FILE_PATH = ms3_options_files/ms3_options_OCO2_example.txt
CLOUDSCREEN_LOG_FILE_NAME_OUT = ./not_used
SOUNDING_LIST_FILE_PATH =

#
# Photon Pathlengths Options
#
ENABLE_PHOTON_PATH_LENGTH_CLOUD_SCREEN = 0
ENABLE_SHADOW = 0
PHOTON_PATH_LENGTH_DIR_PATH = ../dat/photon_path_pdf/
ABS_COEF_FILE_PATH =
ABS_COEF_FILE_PATH =
ABS_COEF_FILE_PATH =

#
# Surface Reflectance Screen Options (Using MODIS MCD43B1)
#
ENABLE_SURFACE_REFLECTANCE = 0
MODIS_BRDF_DIR_PATH = /path/MODIS/MCD43B1/
SURFACE_REFLECTANCE_CONFIG_FILE = ../SURFACE_REFLECTANCE_CONFIG.MEN

#
# Specification File Names
#
PreprocessorSpecificationFileName = /dat_share/hdf/specification_preprocessor_oco2_20131230.dat
L1BSpecificationFileName = /dat_share/hdf/specification_l1b_oco_20140307.dat
MeteorologySpecificationFileName = /dat_share/hdf/specification_met_oco2_sdos_20130430.dat
```

A.1.2 aband_options_oco2.dat

```
#
# Options file for the O2A-Band Cloud Screening
# Blank lines and lines beginning with # are ignored.
# All omitted fields are filled with default values in the code,
# so there is no penalty to omit any fields.
#

# Name of file that contains the pressure levels that retrieval is performed on.
# If omitted, there is a default set of 12 that are used.

# Name of File that contains the SNR-Chi_Squared relationship
CHISQTHRESHOLDSFILE = OCO2_files/simulations/chisq_thresholds.txt

# Name of File that contains the SZA-Surface_Pressure_Offset relationship
PSURFOFFSETFILE = OCO2_files/simulations/psurf_offset_baseline.dat

# Name of file that contains the solar lines to mask out.
# There is a default set of twenty lines used if this file is omitted.

# WHICH SV ELEMENTS TO FIT?
# 1=Psurf; 2=T offset; 3,4=Albedo 1,2; 5=ZLO; 6=DISP MULTIPLIER
SV_SUBSET = 1, 2, 3, 4, 6

# Mask Solar Lines?
MASKSOLARLINES = F

# Solar Zenith beyond which no cloud screening is attempted
MAXSOLARZENITH = 85.0

# SNR below which no cloud screening is attempted
MINSNR = 20.0

# Land Fraction below which a scene is classified as water.
LANDFRACTIONTHRESHOLD = 0.2

# Window of the oxygen A-band to use.
BAND1_WINDOW = 13145.0, 13172.0, 13047.0, 13072.0

# Endpoints of band at which to define the surface albedo
# These must be increasing for GOSAT, decreasing for OCO-2
# Correspond to 0.755, 0.785 micron
BAND1_ENDPOINTS = 13245.0332d0, 12738.8535d0

##---
##--- Surface Pressure Thresholds for various scene types
##---

# Land
PSURF_THRESH_LAND = 25.0
```

```
# Ocean, SNR1 THRESHOLD REGION
PSURF_THRESH_OCEAN_SNR1 = 25.0

# Ocean, SNR2 THRESHOLD REGION
PSURF_THRESH_OCEAN_SNR2 = 50.0

# Ocean, SNR3 THRESHOLD REGION
PSURF_THRESH_OCEAN_SNR3 = 100.0

# First Guess for Surface Pressure Offset
PSURF_OFFSET_FIRST_GUESS = 0.0

# Chisq Multiplier for Land and Ocean
CHISQ_THRESH_MULTIPLIER_LAND = 1.4
CHISQ_THRESH_MULTIPLIER_OCEAN = 1.4

# Reference polarization on which to retrieve
# 1 = P, 2 = S, 3= (P+S)/2
REFERENCE_POLARIZATION = 1

# Ignore or Use the Stokes Coefficients for the chosen polarization?
# For GOSAT, sometimes the stokes coefficients are set to garbage
# so it can be beneficial to ignore them.
IGNORE_STOKES = F

# Number of iterations to perform
N_ITERATIONS = 1

# Calculate True Chisquared or use linear approximation?
CALC_TRUE_CHISQ = F

# Other Logicals
LOCALVERBOSE = F
WRITE_SPECTRUM = F
FIXSURFACEPRESSURE = F
FITZEROLEVELOFFSET = F
PROCESSOCEAN_NONGLINT = F
```

A.1.3 ms3_options_oco2.dat

```
1                                                    ! Instrument Type (1=OCO, 2=GOSAT)
00001 /                                              ! ECMWF profile indices to run, terminate with /
1 /                                                  ! bands to do, terminate with /
"/dat_share/MS3/"                                    ! data directory (relative to run directory)
"./"                                                 ! profile directory
"ms3_output/"                                        ! output directory
".dat"                                               ! profile extension (filenames contain 5 digit profile num)
"gasabs_absco_v420_unscaled_ms3_1.dat"               ! example gas abs file name
"solar_merged_transmittance_res4.txt"                ! solar linelist file name
"suppress"                                           ! instrument dispersion file name (doesn't exist means to
take from scene)
"lsi_bounds_new_dk1.txt"                             ! LSI file name for a single band
"suppress"                                           ! ILS table file name for a single band
"os_absco_v4.0.2_1.dat"                              ! optimal sampling file name for a single band
-25.0 0.0                                            ! obs angles (negative means to take from scene)
-25.0 0.0                                            ! solar angles (negative means to take from scene)
-2                                                   ! Default Surface Type (-1=take from scene,0=Lambertian,8=Cox-
Munk)
0.130 0.175 0.236 0.222 0.142 0.129                  ! surface parameters (6 albedos or
Cox-Munk parameters)
F                                                    ! use mean sun-earth distance = 1 AU ?
-2006 09 14 11 09 51 086                             ! epoch: yr mon day hr min sec msec. (for
solar doppler shift)
0 4                                                  ! nstreams lo, hi
1e-3 3e-4                                            ! max_tau lo,hi
T T                                                  ! delta scaling (lo,hi)
0 1                                                  ! polarization correction (0=none, 1=1OS, 2=2OS)? (lo,hi)
T F                                                  ! truncated doubling? (lo,hi)
F T                                                  ! adding? (lo,hi)
F T                                                  ! PS Correction? (lo,hi)
2e-5 2e-4                                            ! frac convergence for soi (lo,hi)
3e-7 3e-5                                            ! frac convergence for fourier loop (lo,hi)
0 0                                                  ! soi layer init method (lo,hi)
F F                                                  ! verbose rt
1                                                    ! Speedup method (0=none, 1=LSI, 2=EOF)
2                                                    ! nsum for radiative transfer
1016 1016 1016                                       ! number of channels per band
-0.0424717 0.07582 0.097318                          ! Gaussian ILS FWHM values in nm (neg-
ative for table)
8.17e-4 10.0 10.0                                    ! ILS Extent
F                                                    ! output hires spectrum? (F/T)
F                                                    ! output lo-res spectrum? (F/T)
F                                                    ! output lsi files? (F/T)
F                                                    ! output solar files? (F/T)
T                                                    ! use solar doppler shift?
F                                                    ! use instrument doppler shift?
0.0                                                  ! FOV-OCO relative velocity (m/s). +=approaching.
T                                                    ! convolve with solar spectrum?
T                                                    ! use optimal sampling?
3                                                    ! # gas sublayers
F                                                    ! verbose driver output
F                                                    ! Cloud-Free RT?
2                                                    ! Solar Continuum Model
-1.0                                                 ! Surface Pressure [Pa]. Negative means to take from scene.
```

```
0.0                     ! Temperature Offset.
1.0                     ! Water Vapor Scale Factor
1.0                     ! CO2 scale factor
1.0                     ! ILS scale factor
```

A.2 GOSAT retrieval mode

A.2.1 CONFIG.MEN

```
#
# Configuration file for the preprocessor.
# Blank lines and lines beginning with # are ignored.
#
L1B_FILE_PATH = /path1/path2/acos_L1b_example.h5
MET_FILE_PATH = /path1/path2/acos_Ecm_example.h5
QUICK_LOOK_FILE_NAME_OUT =/path1/path2/acos_Cld_example.h5
SUCCESS_SEMAPHORE_FILE_NAME_OUT = ./semaphore.txt

#
# A-Band CloudScreen Options
#
ENABLE_ABO2_CLOUD_SCREEN = 1
ABO2_OPTIONS_FILE_PATH = aband_options_files/aband_options_GOSAT_example.dat
MS3_OPTIONS_FILE_PATH = ms3_options_files/ms3_options_GOSAT_example.txt
CLOUDSCREEN_LOG_FILE_NAME_OUT = ./not_used
SOUNDING_LIST_FILE_PATH =

#
# Photon Pathlengths Options
#
ENABLE_PHOTON_PATH_LENGTH_CLOUD_SCREEN = 0
ENABLE_SHADOW = 0
PHOTON_PATH_LENGTH_DIR_PATH = ../dat/photon_path_pdf/
ABS_COEF_FILE_PATH =
ABS_COEF_FILE_PATH =
ABS_COEF_FILE_PATH =

#
# Surface Reflectance Screen Options (Using MODIS MCD43B1)
#
ENABLE_SURFACE_REFLECTANCE = 0
MODIS_BRDF_DIR_PATH = /path/MODIS/MCD43B1/
SURFACE_REFLECTANCE_CONFIG_FILE = ../SURFACE_REFLECTANCE_CONFIG.MEN

#
# Specification File Names
#
PreprocessorSpecificationFileName = /dat_share/hdf/specification_preprocessor_gosat_20140203.dat
L1BSpecificationFileName = /dat_share/hdf/specification_l1b_gosat_sdos.dat
MeteorologySpecificationFileName = /dat_share/hdf/specification_meteorology_gosat.dat
```

References

Aben, I., O. Hasekamp, and W. Hartmann, 2007: Uncertainties in the space-based measurements of CO_2 columns due to scattering in the Earth's atmosphere. *J. Quant. Spectrosc. Radiat. Transfer*, **104**, 450–459.

Ackerman, S., et al., 2006: Discriminating clear–sky from cloud with MODIS. Algorithm Theoretical Basis Document v5.0, MODIS cloud mask team. Avaialable on-line at http://modis-atmos.gsfc.nasa.gov/_docs/.

Ackerman, S. A., K. I. Strabala, W. P. Menzel, R. A. Frey, C. C. Moeller, and L. E. Gumley, 1998: Discriminating clear sky from clouds with MODIS. *J. Geophys. Res.*, **103 (D24)**, D07206, doi:10.1029/.

Bodhaine, B. A., N. B. Wood, E. G. Dutton, and J. R. Slusser, 1999: On Rayleigh optical depth calculations. *J. Atmos. Oceanic Technol.*, **16**, 1854–1861.

Cox, C. and W. H. Munk, 1954a: The measurement of the roughness of the sea surface from photographs of the sun's glitter. *J. Opt. Soc. Amer.*, **44**, 838–850.

Cox, C. and W. H. Munk, 1954b: Statistics of the sea surface derived from sun glitter. *J. Mar. Res.*, **13**, 198–227.

Crisp, D. and C. Johnson, 2005: The Orbiting Carbon Observatory mission. *J. Atmos. Sci.*, **56**, 193–197, doi:10.1016/j.actaastro.2004.09.032.

Crisp, D., et al., 2004: The Orbiting Carbon Observatory (OCO) mission. *Adv. Space Res.*, **34**, 700–709, doi:10.1016/j.asr.2003.08.062.

Crisp, D., et al., 2012: The ACOS CO_2 retrieval algorithm–part II: Global X_{CO_2} data characterization. *Atmos. Meas. Tech.*, **5**, 687–707, doi:10.5194/amt-5-687-2012.

Day, J. O., C. W. O'Dell, R. Pollock, C. J. Bruegge, D. Rider, D. Crisp, and C. E. Miller, 2011: Preflight spectral calibration of the Orbiting Carbon Observatory. *IEEE Trans. Geosci. Remote Sens.*, **49**, 2793–2801, doi:10.1109/TGRS.2011.2107745.

Frankenberg, C., U. Platt, and T. Wagner, 2005: Iterative maximum a posteriori (IMAP)-DOAS for retrieval of strongly absorbing trace gases: Model studies for $CH_4$and CO_2retrieval from near infrared spectra of SCIAMACHY onboard ENVISAT. *Atmos. Chem. Phys.*, **5**, 9–22.

Frey, R. A., S. A. Ackerman, Y. Liu, K. I. Strabala, H. Zhang, J. R. Key, and X. Wang, 2008: Cloud detection with MODIS. Part I: Improvements in the MODIS cloud mask for collection 5. *J. Atmos. Oceanic Technol.*, **25**, 1057–1072, doi:10.1175/2008JTECHA1052.1.

Gao, B., A. F. Goetz, and W. J. Wiscombe, 1993: Cirrus cloud detection from airborne imaging spectromter data using the 1.38 μm water vapor band. *Geophys. Res. Lett.*, **20**, 301–304.

Gao, B., P. Yang, W. Han, R.-R. Li, and W. Wiscombe, 2002: An algorithm using visible and 1.375-μm channels to retrieve cirrus cloud reflectances from aircraft and satellite data. *IEEE Trans. Geosci. Remote Sens.*, **40**, 1659–1688, doi:10.11.

Heidinger, A. K., C. O'Dell, R. Bennartz, and T. Greenwald, 2006: The successive-order-of-interaction radiative transfer model: Part I: Model development. *J. Clim. Appl. Meteorol.*, **45**, 1388–1402.

Ishida, H. and T. Y. Nakajima, 2009: Development of an unbiased cloud detection algorithm for a spaceborne multispectral imager. *J. Geophys. Res.*, **114 (D07206)**, doi: 10.1029/2008JD010710.

Mandrake, L., C. Frankenberg, C. O'Dell, G. Osterman, P. Wennberg, and D. Wunch, 2013: Semi-autonomous sounding selection for OCO-2. *Atmos. Meas. Tech.*, **6**, 2851–2864, doi: 10.5194/amt-6-2851-2013.

Meyer, K., P. Yang, and B. Gao, 2007: Ice cloud optical depth from MODIS cirrus reflectance. *IEEE Trans. Geosci. Remote Sens.*, **4 (3)**, 471–474, doi: 10.1109/LGRS.2007.897428.

O'Brien, D. M., I. Polonsky, C. O'Dell, and A. Carheden, 2009: Orbiting Carbon Observatory (OCO), algorithm theoretical basis document: The OCO simulator. Technical report ISSN 0737-5352-85, Cooperative Institute for Research in the Atmosphere, Colorado State University, 1-48 pp.

O'Brien, D. M. and P. J. Rayner, 2002: Global observations of the carbon budget 2. CO_2 column from differential absorption of reflected sunlight in the 1.61 μm band of CO_2. *J. Geophys. Res.*, **107 (D18)**, ACH 6-1, doi:10.1029/2001JD000617.

O'Dell, C., A. K. Heidinger, T. Greenwald, P. Bauer, and R. Bennartz, 2006: The successive-order-of-interaction radiative transfer model: Part II: Model performance and applications. *J. Clim. Appl. Meteorol.*, **45**, 1403–1413.

O'Dell, C. W., 2010: Acceleration of multiple-scattering, hyperspectral radiative transfer calculations via low-streams interpolation. *J. Geophys. Res.*, **115 (D10206)**, doi: 10.1029/2009JD012803.

O'Dell, C. W., et al., 2012: The ACOS CO_2 retrieval algorithm–Part 1: Description and validation against synthetic observations. *Atmos. Meas. Tech.*, **5**, 99–121, doi:10.5194/amt-5-99-2012.

Rodgers, C. D., 2000: *Inverse Methods For Atmospheric Sounding: Theory and Practice*. World Scientific Publishing Co. Pte. Ltd.

Salstein, D. A., R. M. Ponte, and K. Cady-Pereira, 2008: Uncertainties in atmospheric surface pressure fields from global analyses. *J. Geophys. Res.*, **113 (D14)**, doi: 10.1029/2007JD009531.

Stephens, G. L. and A. Heidinger, 2000: Molecular line absorption in a scattering atmosphere. part I: Theory. *J. Atmos. Sci.*, **57**, 1599–1614.

Taylor, T. E., et al., 2012: Comparison of cloud-screening methods applied to GOSAT near-infrared spectra. *IEEE Trans. Geosci. Remote Sens.*, **50 (1)**, 295–309, doi: 10.1109/TGRS.2011.2160270.